ELSEVIER

Decision Support Systems 30 (2000) 171–186

Decision Support Systems

www.elsevier.com/locate/dsw

Exploring the use of concept spaces to improve medical information retrieval

Andrea L. Houston[a,*], Hsinchun Chen[b,1], Bruce R. Schatz[c,2], Susan M. Hubbard[d,3], Robin R. Sewell[e,4], Tobun D. Ng[f,5]

[a] ISDS Department, College of Business Administration, Louisiana State University, 3178B4 CEBA, Baton Rouge, LA 70803, USA
[b] Management Information Systems Department, University of Arizona, Tucson, AZ 85721, USA
[c] Community Architecture for Network Information Systems (CANIS) Laboratory, University of Illinois, Urbana–Champaign, Urbana, IL 61820, USA
[d] Director of the International Cancer Information Center, National Cancer Institute, Bethesda, MD 20852, USA
[e] School of Library Science, University of Arizona, Tucson, AZ 85721, USA
[f] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

This research investigated the application of techniques successfully used in previous information retrieval research, to the more challenging area of medical informatics. It was performed on a biomedical document collection testbed, CANCERLIT, provided by the National Cancer Institute (NCI), which contains information on all types of cancer therapy. The quality or usefulness of terms suggested by three different thesauri, one based on MeSH terms, one based solely on terms from the document collection, and one based on the Unified Medical Language System (UMLS) Metathesaurus, was explored with the ultimate goal of improving CANCERLIT information search and retrieval.

Researchers affiliated with the University of Arizona Cancer Center evaluated lists of related terms suggested by different thesauri for 12 different directed searches in the CANCERLIT testbed. The preliminary results indicated that among the thesauri, there were no statistically significant differences in either term recall or precision. Surprisingly, there was almost no overlap of relevant terms suggested by the different thesauri for a given search. This suggests that recall could be significantly improved by using a combined thesaurus approach. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Information retrieval; Medical informatics; Medical information retrieval; Concept space; MeSH terms; UMLS Metathesaurus

1. Introduction

Medicine is a dynamic field incorporating numerous specialties, each with its own preferred terminology. This diversity of vocabularies can be an obstacle for medical professionals requiring access to current medical information [16]. While advances in medical database technology have improved information accessibility, retrieval speed, and searching flexibility, they have not resolved the problems of

[*] Corresponding author. Tel.: +1-225-578-2503; fax: +1-225-578-2511.

E-mail addresses: ahoust2@lsu.edu (A.L. Houston), hchen@bpa.arizona.edu (H. Chen), schatz@canis.uiuc.edu (B.R. Schatz), sh68q@nih.gov (S.M. Hubbard), rrs@ai2.bpa.arizona.edu (R.R. Sewell), dorbin.ng@cs.cmu.edu (T.D. Ng).

[1] Tel.: +1-520-621-4153.

[2] Tel.: +1-217-244-0651.

[3] Tel.: +1-301-496-9096.

[4] Tel.: +1-520-621-2748.

[5] Tel.: +1-412-268-4499.

0167-9236/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-9236(00)00097-X

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2000**	2. REPORT TYPE	3. DATES COVERED **00-00-2000 to 00-00-2000**

4. TITLE AND SUBTITLE
Exploring the use of concept spaces to improve medical information retrieval

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Arizona ,Management Information Systems Department ,Tucson,AZ,85721

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research investigated the application of techniques successfully used in previous information retrieval research, to the more challenging area of medical informatics. It was performed on a biomedical document collection testbed CANCERLIT, provided by the National Cancer Institute NCI., which contains information on all types of cancer therapy. The quality or usefulness of terms suggested by three different thesauri, one based on MeSH terms, one based solely on terms from the document collection, and one based on the Unified Medical Language System UMLS.Metathesaurus, was explored with the ultimate goal of improving CANCERLIT information search and retrieval. Researchers affiliated with the University of Arizona Cancer Center evaluated lists of related terms suggested by different thesauri for 12 different directed searches in the CANCERLIT testbed. The preliminary results indicated that among the thesauri, there were no statistically significant differences in either term recall or precision. Surprisingly, there was almost no overlap of relevant terms suggested by the different thesauri for a given search. This suggests that recall could be significantly improved by using a combined thesaurus approach.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**16**	

vocabulary differences among biomedical specialties, variations in indexing and classification systems, nor variations in information accessing systems. Medical information retrieval, as a specialized case of information retrieval, is subject to classic information retrieval problems such as: “information overload” [3], the “vocabulary problem” (semantic barrier [17]), synonymy and polysemy. It also has some interesting problems of its own [12,14,15].

The community of medical information users is extremely varied in its level of biomedical expertise, its familiarity with various biomedical indexing vocabularies and its information usage requirements. For example, biomedical expertise ranges from patients and families encountering terms for the first time, to specialists in focused research areas who are considered experts. Compounding this problem is the fact that there is no single commonly accepted biomedical indexing vocabulary. This lack of an information standard and the existence of thousands of different medical databases containing information that can be formatted, indexed and stored in a variety of different ways make it difficult, if not impossible, to locate and exchange medical information. Users requiring information from a variety of medical sources may have to learn several different information retrieval systems and several different indexing vocabularies to locate the information they need.

Depending on medical information usage requirements, the goals of indexing vocabularies may conflict. For example, biomedical research information, databases of clinical studies or drug trials, and medical insurance databases all need to have data organized or summarized by categories (*generalization*). However, primary care professionals dealing with individual patient records require a detailed, precise and expressive vocabulary that can accurately describe patient information [11,13]. Patient records can be a composite of every potential data format (numeric, free text, tables, graphs, images and audio). Patient record information systems, therefore, require a standard vocabulary that can *specialize* (the direct opposite of generalization) and accommodate a massive quantity of highly variable and volatile information, thereby increasing medical information system challenges.

We are investigating improving medical information retrieval by building on techniques successfully applied to other information retrieval domains (e.g. Worm/Fly Genome, the Internet, and a large scientific abstract collection). Previous research demonstrated that the creation of automatically generated concept spaces (thesauri) is an efficient, effective technique to improve document precision and recall in *directed* searches of large information spaces. The Worm/Fly genome research [5] indicated that a combined thesaurus approach could improve recall without sacrificing precision. Currently, we are investigating augmenting automatically generated concept space terms with terms from existing medical thesauri: Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) Metathesaurus, both developed and maintained by the National Library of Medicine (NLM).

2. Literature review

There are four major approaches to textual medical information retrieval. They are: keyword indexing and retrieval (traditional method), statistically based methods (Salton-based syntactic techniques), relevance feedback (using searcher feedback to improve future searches), and semantic methods (including extensions to Salton’s techniques and Natural Language Processing).

2.1. Human indexing and keyword search

The most common example of the keyword approach is human indexing (using a standard set of pre-determined subject terms that domain experts assign to documents). This approach relies entirely on indexer expertise in both subject domain and standard indexing vocabularies. Previous research [2] demonstrated that different, well-trained indexers often assign different terms to the same document (*synonymy*) and that an indexer may use different terms for the same document at different times. Meanwhile, different users tend to use different terms to seek identical information (*polysemy*). Furnas et al. [10], showed that the probability of two people using the same term to classify an object was less than 20%.

Because of these discrepancies, an exact match between searcher terms and indexer terms is unlikely, resulting in poor document recall and precision. Individual keywords are simply not adequate discriminators of semantic content. Furthermore, manual indexing is too time-consuming for processing large volumes of information. Human indexing methods need to be either supplemented or replaced by more effective and efficient techniques, hence, the major research effort in developing automatic indexing techniques.

2.2. Statistical techniques — vector space and thesaurus

Other approaches to addressing the vocabulary problem include using either a thesaurus or a vector space document representation [19]. Thesauri (mainly used to expand users' queries by translating query terms into alternative terms that match document indexes) can be generated manually or automatically. Srinivasan [21] presents a medical informatics example that evaluates different query expansion strategies using a MEDLINE testbed. Most automatically generated thesauri are syntactically based techniques that use vector space document representation and word statistical co-occurrence analysis. Many incorporate other statistical techniques such as cluster analysis, co-occurrence efficiency analysis, and factor analysis. Relationships are then represented in mathematical matrices.

Most statistical methods concentrate on solving synonymy by adding associative terms to keyword indexes. A major disadvantage is that some added terms have meanings that are different from the intended meanings, resulting in rapid degradation of information retrieval precision. Cimino et al. [7] has an interesting discussion on the problems of automated medical information translation using thesauri. Automatically generated neural-like thesauri (concept spaces) provide an alternative to traditional thesauri. In a neural knowledge base, concepts (terms) are represented as nodes, and relationships as weighted links. The associative memory feature of this thesaurus type allows a new paradigm for knowledge discovery and document searching using spreading activation algorithms (e.g. Hopfield net).

2.3. Relevance feedback

Relevance feedback is a method that automates the intellectual process of evaluating the results of an initial search to improve future searches. It can be used with both vector queries and Boolean searches. Salton and Buckley [20] discuss a variety of procedures for relevance feedback and outline three benefits: automatic expansion of queries, gradual advancement towards the subject, and selective key term emphasis. Their experiments demonstrated that relevance feedback can be very effective (90% precision [20]) and its usefulness has been well-documented in TIPSTER conferences. The disadvantages of this technique include the facts that concerns over processing speed and information storage outweigh the benefits, and that it cannot improve effective queries.

2.4. Semantic approaches

Another approach is to index documents semantically, allowing users to search using conceptual *meanings* instead of keywords. Multi-dimensional semantic space techniques attempt to enhance information retrieval by placing documents (vectors) by meaning in a designated space. The most representative multi-dimensional semantic space techniques are Metric Similarity Modeling (MSM) and Latent Semantic Indexing (LSI).

MSM represents both queries and documents with vectors in a multi-dimensional semantic space using techniques from Multi-Dimensional Scaling [1]. Document vectors are computed using standard statistical techniques and then placed in a multi-dimensional semantic space, their positions determined by similarity constraints. One disadvantage of MSM is that it can only be used when external sources exist to determine similarity constraints (i.e. co-citation analysis, relevance feedback or document classification information — Library of Congress Subject Headings or Compendex Classification Codes).

LSI [8] is an optimal method of MSM. It represents documents, queries and terms as vectors in a matrix determined by multi-dimensional Singular Value Decomposition. LSI takes advantage of the fact that semantic relations exist within a document and attempts to place similar documents close to each other in a multi-dimensional space. Chute et al. [6] have applied the technique to medical informatics. Deerwester et al. [8] have tested LSI on two standard document collections (MED and CISI — chosen because relevance judgments already existed) with promising results in both document recall and precision. LSI proved equal to or better than either simple term matching or SMART, and better than Voorhees' term disambiguation process. Unfortunately, the meanings of these mathematically derived semantic techniques are difficult to understand and computationally cumbersome. Their usefulness in suggesting meaningful indexing or searching terms has not been validated on a large real-world collection.

2.5. UMLS

In 1986, NLM began developing the UMLS to address medical vocabulary problems by "improving the ability of computer programs to 'understand' biomedical meaning in user inquiries and then using this understanding to retrieve and integrate relevant machine-readable information" [16]. The UMLS has four components: the Metathesaurus, the Specialist Lexicon, the Semantic Net, and the Information Sources Map. The Metathesaurus is the largest and most complex component, incorporating 589 000 names for 253 000 concepts from more than 30 biomedical vocabularies, thesauri, and classification systems (including MeSH, SNOMED, COSTAR, ICD-9CM, and Dorland's Illustrated Medical Dictionary, 27th edn.). The Metathesaurus is not intended to serve as an "overarching classification system" or controlled vocabulary, but to facilitate translation and interpretation of biomedical terminology across vocabularies.

NLM makes UMLS copies available to researchers for the development of biomedically related expert systems, automatic indexing and classification tools, and tools to index patient records. Research in this area includes the SAPHIRE project (Oregon Health Sciences University), the Internet Grateful Med interface for MEDLARS databases (COACH Browser), the Interactive Query Workstation (IQW), the InterMed Vocabulary Server project (Stanford, Columbia, Harvard and the University of Utah), and the Group Health Cooperative of Puget Sound. Metathesaurus use has been shown to significantly increase (60–88%) document recall rate (over MeSH terms) [18].

3. CANCERLIT experiment

CANCERLIT contains bibliographic records (predominantly abstracts) from biomedical journals on research related to cancer biology, etiology, screening, prevention, and treatment published between 1963 and today. Approximately 200 core journals account for the majority of the collection. Additional citations come from journals, scientific meeting proceedings, books, dissertations, technical reports, and other publications. The National Cancer Institute (NCI) and NLM share processing costs, therefore, many CANCERLIT citations are cross-indexed in MEDLINE. CANCERLIT is updated monthly to ensure that the most current published cancer research results are available (see acknowledgments). There are more than 1.2 million records in the complete collection, which NCI estimates increases annually by approximately 90 000 abstracts. The record format includes the following fields: authors and addresses, MeSH headings indexing the document, and the document's source, title, and abstract. More detailed information is available from NCI.

3.1. CANCERLIT concept space

Our CANCERLIT testbed contains 2 months of CANCERLIT data (May and June 1996). The approximately 10 000 abstracts take up 40 MB of memory, and took roughly 1 h to create on an HP 9000 workstation. We have since expanded the testbed to include the last 5 years of CANCERLIT documents (seeai.bpa.arizona.edu/CancerLit).

Two different options were used to create the prototype CANCERLIT concept space. A MeSH-based CANCERLIT concept space was created using existing MeSH terms that index each document, and an Automatic Indexing CANCERLIT concept space was generated using automatic indexing techniques (word identification, stop wording, stemming, term-phrase formation, and tf * idf term weighing). Concept space (automatic thesaurus) creation is a standard process that can be applied to a variety of different kinds of textual information. Users with different levels of expertise have successfully used the output in different subject domains. Fig. 1 illustrates the complete process as it might be applied to medical knowledge spaces. A brief overview of the process is described below.

3.1.1. Document collection

In any automatic thesaurus building effort, the first task is to identify the collection of documents that will serve as the basis of the thesaurus. We used a 2-month collection of CANCERLIT documents.

3.1.2. Automatic indexing

The purpose of this step is to *automatically* identify each document's content. We used a Salton-based

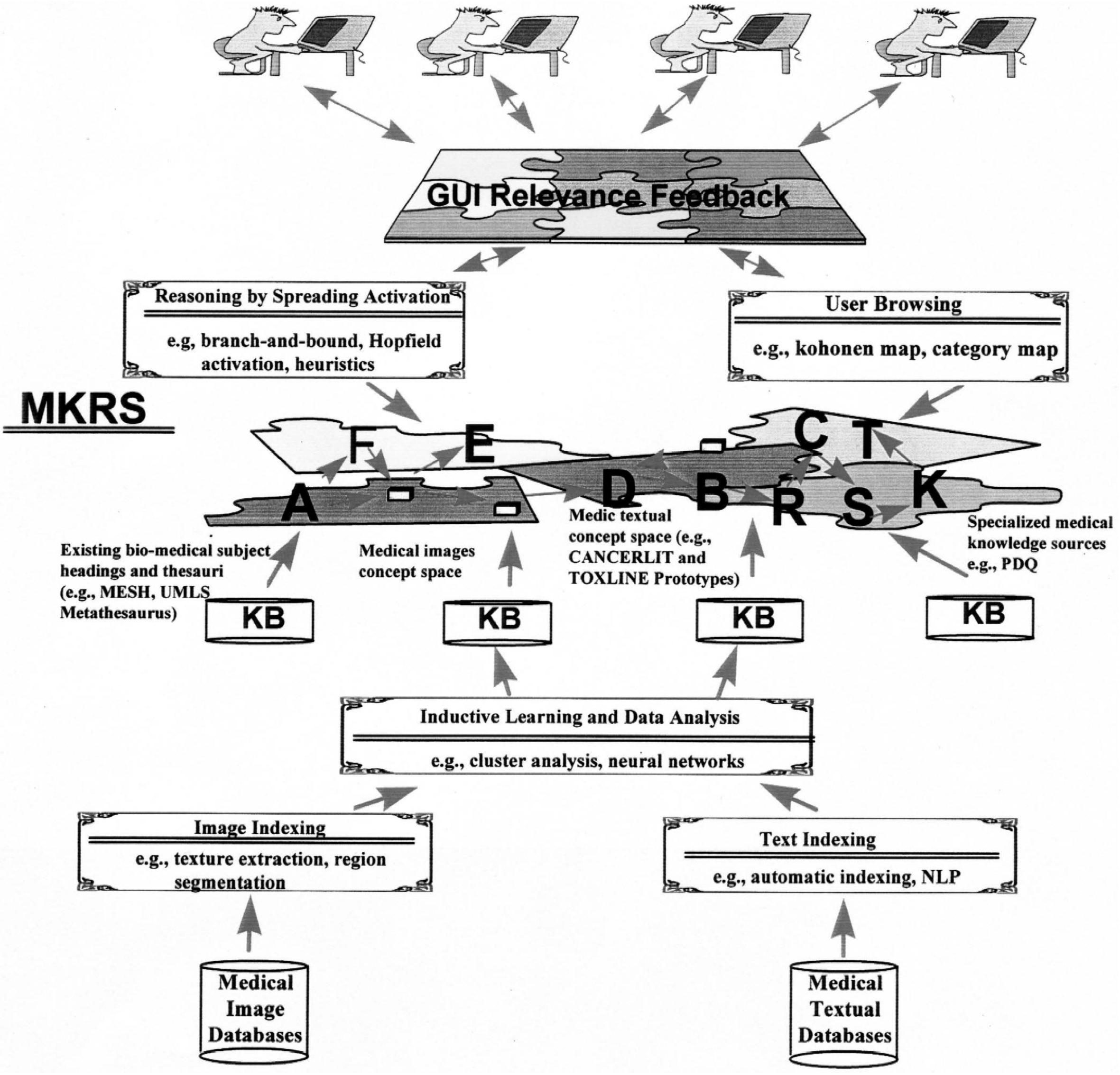

GUI Relevance Feedback
Reasoning by Spreading Activation
e.g, branch-and-bound, Hopfield activation, heuristics
User Browsing
e.g., kohonen map, category map
MKRS
F
E
C
T
A
D
B
R
S
K
Existing bio-medical subject headings and thesauri (e.g., MESH, UMLS Metathesaurus)
Medical images concept space
Medic textual concept space (e.g., CANCERLIT and TOXLINE Prototypes)
Specialized medical knowledge sources e.g., PDQ
KB
KB
KB
KB
Inductive Learning and Data Analysis
e.g., cluster analysis, neural networks
Image Indexing
e.g., texture extraction, region segmentation
Text Indexing
e.g., automatic indexing, NLP
Medical Image Databases
Medical Textual Databases

Fig. 1. The Medical Knowledge Representation System (MKRS) Architecture.

technique [19] that identifies document subject descriptors and computes descriptor frequency for the entire collection. Then, a stop-word list eliminates non-content bearing words (e.g. "the", "a", "on", "in"), and a stemming algorithm identifies the remaining word stems. A document term-frequency requirement removes "noise".

3.1.3. Co-occurrence analysis

The importance of each descriptor (term) in representing document content varies. Using term frequency and inverse document frequency, the cluster analysis step assigns weights to each document term to represent term importance. Term frequency indicates how often a particular term occurs in the *entire* collection. Inverse document frequency (indicating term specificity) allows terms to have different strengths (importance) based on specificity. A term can be a one-, two-, or three-word phrase. Fig. 2 describes the frequency computation. Cluster analysis is then used to convert raw data indexes and weights into a matrix indicating term similarity/dissimilarity using a distance computation based on Chen and Lynch's [4] *asymmetric* "Cluster Function" which represents term association better than the cosine function (see Fig. 3 for more detail). A net-like concept space of terms and weighted relationships is then created, using the cluster function.

3.1.4. Associative retrieval

The Hopfield algorithm is ideal for concept-based information retrieval. Each term in the network-like thesaurus is treated as a neuron and the asymmetric weight between any two terms is the unidirectional, weighted connection between neurons. With user-supplied terms as input patterns, the Hopfield algorithm activates term neighbors (strongly associated terms), combines weights from all associated neighbors (by adding collective association strengths), and repeats this process until convergence (see Fig. 4).

Fig. 5 illustrates a CANCERLIT session. The user entered the term "breast cancer" and the prototype's combined MeSH/Automatic Index concept space suggests a list of related terms. Next, the user selects the term "Family History" from the list of related terms, which combines it with the original term "breast cancer" to narrow the search. The prototype finds 506 documents that contain the two terms and the user selects the first document to review.

3.2. Experimental design

The primary goal in this preliminary phase was to evaluate the usefulness of suggested terms from three different thesauri:

- MeSH concept space — a thesaurus based on NLM's controlled medical information retrieval controlled vocabulary: MeSH.
- Auto Index concept space — our *automatically* generated thesaurus based exclusively on terms contained in the collection's documents.
- Internet Grateful Med — the most commonly cited on-line tool based on the UMLS Metathesaurus.

The combined weight of term j in document i, d_{ij} was computed as follows:

$$d_{ij} = tf_{ij} \text{ x } \log[(N/df_j) \text{ x } w_j]$$

where N represents the total number of documents in the collection, tf_{ij} represents the number of occurrences of term j in document i, w_j represents the number of words in descriptor T_j, and df_j, represents the number of document in a collection of n documents in which term j occurs. Multiple-word terms were assigned heavier weights as they usually convey more precise semantic meaning.

Fig. 2. Frequency computation.

$$\text{Cluster Weight}(T_j, T_k) = [\Sigma^{n}{}_{i=1} (d_{ijk}) / \Sigma^{n}{}_{i=1} (d_{ij})] \text{ x } \text{Weighting Factor}(T_k)$$

$$\text{Cluster Weight}(T_k, T_j) = [\Sigma^{n}{}_{i=1} (d_{ikj}) / \Sigma^{n}{}_{i=1} (d_{ik})] \text{ x } \text{Weighting Factor}(T_j)$$

These two equations indicate the similarity weights from term T_j to term T_k (the first equation) and from term T_k to term T_j (the second equation). d_{ij} and d_{ik} are frequency computations from the previous step. d_{ijk} represents the combined weight of both descriptors T_j and T_k in document i defined as:

$$d_{ijk} = tf_{ijk} \text{ x } \log[(N/df_{jk}) \text{ x } w_j]$$

Co-occurrence analysis *penalizes* general terms using the following weights (similar to the *inverse document frequency* function) allowing the thesaurus to make more precise suggestions:

$$\text{Weighting Factor}(T_k) = \log (N/df_k) / \log (N)$$

$$\text{Weighting Factor}(T_j) = \log (N/df_j) / \log (N)$$

Fig. 3. Cluster analysis computations.

Our subjects were five cancer researchers affiliated with the University of Arizona Cancer Center, and a veterinarian. Phase one of the experiment involved evaluating term usefulness during a *directed* search of the CANCERLIT testbed. Twelve searches were performed on each of the three thesauri (36 total searches). First, we demonstrated each thesaurus using a subject-provided term. Then, each subject was asked to provide one or two terms to begin a document search, and to suggest five related terms for each search term.

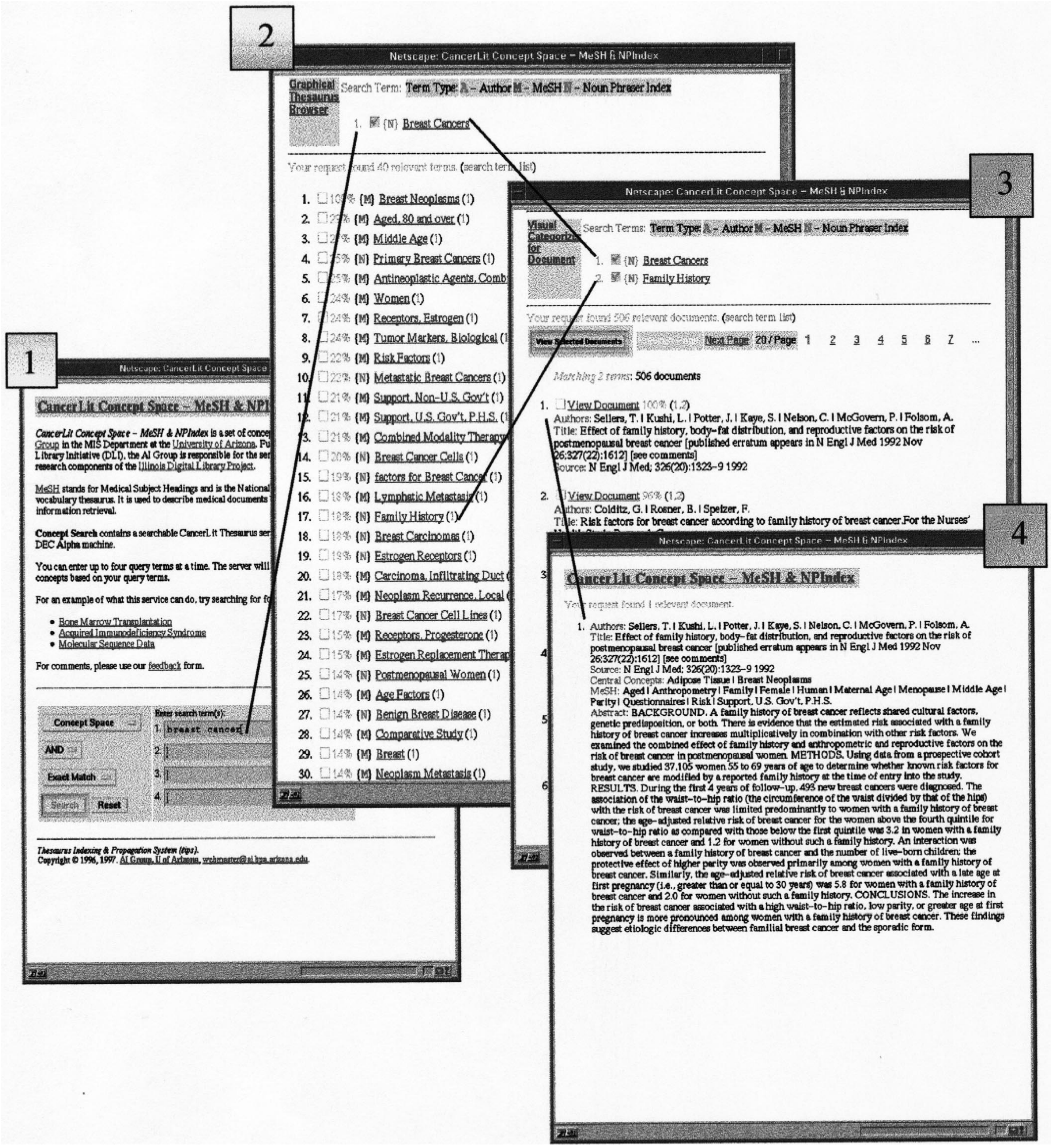
During phase two, a subject-supplied search term was entered into a thesaurus and subjects evaluated the top 40 thesaurus-suggested terms ("relevant" or "not relevant"). This step was repeated for the other two thesauri. The entire process was then repeated for other subject-supplied term(s). The order in which we searched the thesauri was pre-assigned by subject

The Hopfield net algorithm uses an iterative activation process:

$$\mu_j(t+1) = f_s[\Sigma^{n-1}{}_{i=0}\, t_{ij}\mu_i(t)],\ 0 \leq j \leq n\text{-}1$$

$\mu_j(t+1)$ is the activation value of term j at iteration t+1, t_{ij} is the co-occurrence weight from term i to term j, and f_s is the continuous SIGMOID transformation function (normalizes any value to between 0 and 1). This formula shows the *parallel relaxation* property of the Hopfield net.

Fig. 4. Hopfield net algorithm.



Netscape: CancerLit Concept Space

CancerLit Concept Space – MeSH & NPI

CancerLit Concept Space – MeSH & NPIndex is a set of concep
Group in the MIS Department at the University of Arizona. Fu
Library Initiative (DLI), the AI Group is responsible for the ser
research components of the Illinois Digital Library Project.

MeSH stands for Medical Subject Headings and is the National
vocabulary thesaurus. It is used to describe medical documents
information retrieval.

Concept Search contains a searchable CancerLit Thesaurus ser
DEC Alpha machine.

You can enter up to four query terms at a time. The server will
concepts based on your query terms.

For an example of what this service can do, try searching for fo

- Bone Marrow Transplantation
- Acquired Immunodeficiency Syndrome
- Molecular Sequence Data

For comments, please use our feedback form.

Concept Space | AND | Exact Match | Search | Reset

Enter search term(s):
1. breast cancer
2.
3.
4.

Thesaurus Indexing & Propagation System (tips).
Copyright © 1996, 1997. AI Group, U of Arizona, webmaster@ai.bpa.arizona.edu

Netscape: CancerLit Concept Space – MeSH & NPIndex

Graphical Thesaurus Browser

Search Term: Term Type: A – Author M – MeSH N – Noun Phraser Index

1. ☑ {N} Breast Cancers

Your request found 40 relevant terms. (search term list)

1. □ 100% {M} Breast Neoplasms (1)
2. □ 29% {M} Aged, 80 and over (1)
3. □ 27% {M} Middle Age (1)
4. □ 25% {N} Primary Breast Cancers (1)
5. □ 25% {M} Antineoplastic Agents, Comb
6. □ 24% {M} Women (1)
7. □ 24% {M} Receptors, Estrogen (1)
8. □ 24% {M} Tumor Markers, Biological (1
9. □ 22% {M} Risk Factors (1)
10. □ 22% {N} Metastatic Breast Cancers (1)
11. □ 21% {M} Support, Non-U.S. Gov't (1)
12. □ 21% {M} Support, U.S. Gov't, P.H.S. (1
13. □ 21% {M} Combined Modality Therapy
14. □ 20% {N} Breast Cancer Cells (1)
15. □ 19% {N} factors for Breast Cancer (1)
16. □ 18% {M} Lymphatic Metastasis (1)
17. □ 18% {N} Family History (1)
18. □ 18% {N} Breast Carcinomas (1)
19. □ 19% {N} Estrogen Receptors (1)
20. □ 18% {M} Carcinoma, Infiltrating Duct (
21. □ 17% {M} Neoplasm Recurrence, Local (
22. □ 17% {N} Breast Cancer Cell Lines (1)
23. □ 15% {M} Receptors, Progesterone (1)
24. □ 15% {M} Estrogen Replacement Therap
25. □ 14% {N} Postmenopausal Women (1)
26. □ 14% {M} Age Factors (1)
27. □ 14% {N} Benign Breast Disease (1)
28. □ 14% {M} Comparative Study (1)
29. □ 14% {M} Breast (1)
30. □ 14% {M} Neoplasm Metastasis (1)

Netscape: CancerLit Concept Space – MeSH & NPIndex

Visual Categorizer for Document

Search Terms: Term Type: A – Author M – MeSH N – Noun Phraser Index

1. ☑ {N} Breast Cancers
2. ☑ {N} Family History

Your request found 506 relevant documents. (search term list)

View Selected Documents | Next Page | 20 / Page | 1 2 3 4 5 6 7 ...

Matching 2 terms: 506 documents

1. □ View Document 100% (1,2)
Authors: Sellers, T. | Kushi, L. | Potter, J. | Kaye, S. | Nelson, C. | McGovern, P. | Folsom, A.
Title: Effect of family history, body–fat distribution, and reproductive factors on the risk of postmenopausal breast cancer [published erratum appears in N Engl J Med 1992 Nov 26;327(22):1612] [see comments]
Source: N Engl J Med; 326(20):1323–9 1992

2. □ View Document 96% (1,2)
Authors: Colditz, G. | Rosner, B. | Speizer, F.
Title: Risk factors for breast cancer according to family history of breast cancer.For the Nurses'

Netscape: CancerLit Concept Space – MeSH & NPIndex

CancerLit Concept Space – MeSH & NPIndex

Your request found 1 relevant document.

1. Authors: Sellers, T. | Kushi, L. | Potter, J. | Kaye, S. | Nelson, C. | McGovern, P. | Folsom, A.
Title: Effect of family history, body–fat distribution, and reproductive factors on the risk of postmenopausal breast cancer [published erratum appears in N Engl J Med 1992 Nov 26;327(22):1612] [see comments]
Source: N Engl J Med; 326(20):1323–9 1992
Central Concepts: Adipose Tissue | Breast Neoplasms
MeSH: Aged | Anthropometry | Family | Female | Human | Maternal Age | Menopause | Middle Age | Parity | Questionnaires | Risk | Support, U.S. Gov't, P.H.S.
Abstract: BACKGROUND. A family history of breast cancer reflects shared cultural factors, genetic predisposition, or both. There is evidence that the estimated risk associated with a family history of breast cancer increases multiplicatively in combination with other risk factors. We examined the combined effect of family history and anthropometric and reproductive factors on the risk of breast cancer in postmenopausal women. METHODS. Using data from a prospective cohort study, we studied 37,105 women 55 to 69 years of age to determine whether known risk factors for breast cancer are modified by a reported family history at the time of entry into the study. RESULTS. During the first 4 years of follow–up, 493 new breast cancers were diagnosed. The association of the waist–to–hip ratio (the circumference of the waist divided by that of the hips) with the risk of breast cancer was limited predominantly to women with a family history of breast cancer; the age–adjusted relative risk of breast cancer for the women above the fourth quintile for waist–to–hip ratio as compared with those below the first quintile was 3.2 in women with a family history of breast cancer and 1.2 for women without such a family history. An interaction was observed between a family history of breast cancer and the number of live–born children; the protective effect of higher parity was observed primarily among women with a family history of breast cancer. Similarly, the age–adjusted relative risk of breast cancer associated with a late age at first pregnancy (i.e., greater than or equal to 30 years) was 5.8 for women with a family history of breast cancer and 2.0 for women without such a family history. CONCLUSIONS. The increase in the risk of breast cancer associated with a high waist–to–hip ratio, low parity, or greater age at first pregnancy is more pronounced among women with a family history of breast cancer. These findings suggest etiologic differences between familial breast cancer and the sporadic form.

Fig. 5. CANCERLIT session: User entered "breast cancer" in (1). CANCERLIT suggests related terms in (2). User selects "Family History". System locates 506 documents related to "breast cancer" and "Family History" in (3). User chooses to read first document in (4).

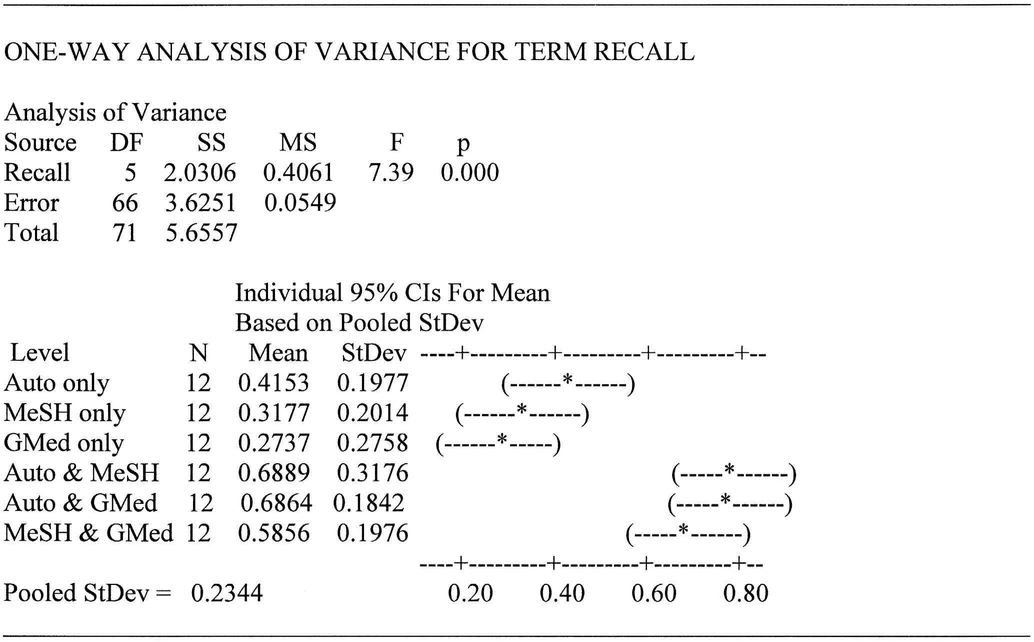
number in a random fashion. We recorded the verbal protocols of the searching and evaluation process. Subjects were allowed to access documents linked to the thesaurus-suggested terms, but we did not evaluate that part of the search. Next, we solicited feedback on the usefulness of the three thesauri, the users' searching experiences with CANCERLIT, MEDLINE, and Internet Grateful Med, and our user interface. Precision and recall for phase two was computed.

3.3. Precision and recall

Precision and recall were calculated as follows: (1) "very relevant" terms — 1 point; (2) "possibly relevant" terms — 0.5 points; and (3) "not relevant" terms — zero points. For each search, all relevant terms from each thesaurus were combined (eliminating duplicates) for a total relevant score. Term recall for each thesaurus was calculated by dividing the relevant score for that thesaurus by the total relevant score for all thesauri. Term precision for each thesaurus was calculated by dividing the total relevant score for the thesaurus by the total number of terms suggested by the thesaurus.

Fig. 6 illustrates Minitab's one-way ANOVA test for term recall for each thesaurus, and for the various thesauri combinations. Fig. 7 illustrates the same information for term precision. There were no statistically significant differences among the three thesauri in term recall or precision, indicating that terms suggested by our tool are comparable to terms suggested by Internet Grateful Med (UMLS Metathesaurus-based) and MeSH indexing terms. Based on subject qualitative feedback, we believe this is partially due to the prototype's size (2 months worth of data vs. the entire MEDLINE collection). We were pleased that our tool could perform at a comparable level based on such limited input.

An interesting result from this research is the lack of duplicate relevant terms suggested by the three different thesauri. In previous research, the most relevant terms typically were suggested by all thesauri. We were surprised at the lack of term overlap (for example, four searches had *no* overlapping terms, and four searches had only *one* overlapping term), which suggests that a combined approach would probably be more useful to searchers. Subjects confirmed this in their verbal feedback, and there is supporting evidence in the literature [5,9,21]. Our data also support the literature's premise that thesaurus combination can increase recall without sacrificing precision.

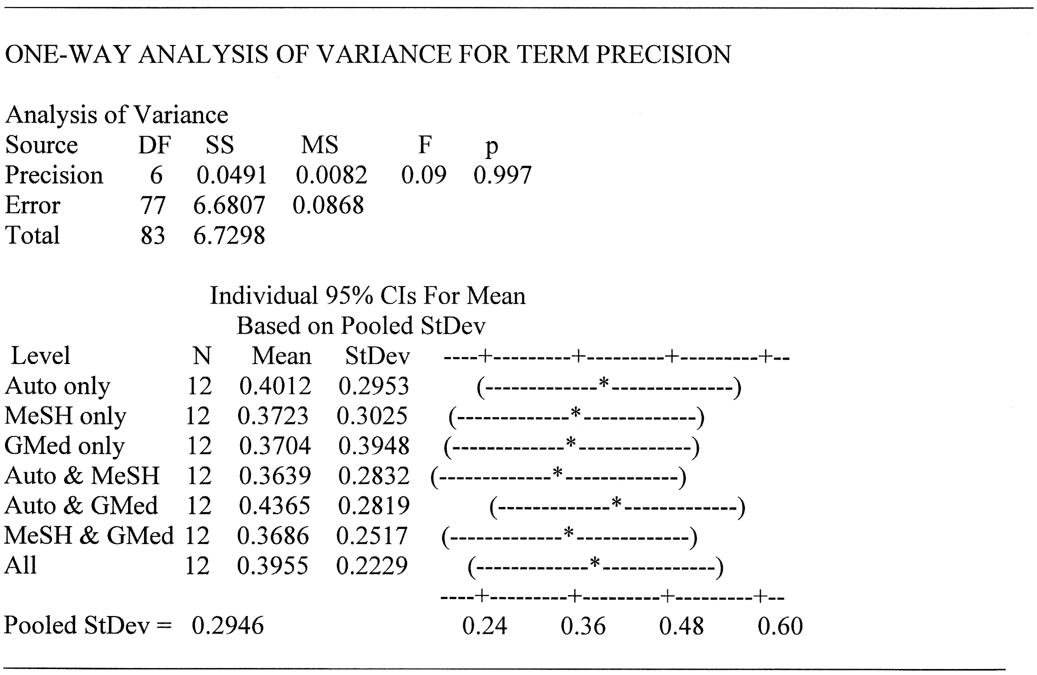


```
ONE-WAY ANALYSIS OF VARIANCE FOR TERM RECALL

Analysis of Variance
Source   DF      SS       MS       F      p
Recall    5   2.0306   0.4061   7.39  0.000
Error    66   3.6251   0.0549
Total    71   5.6557

                        Individual 95% CIs For Mean
                        Based on Pooled StDev
 Level          N    Mean    StDev  ----+---------+---------+---------+--
Auto only      12  0.4153  0.1977          (------*------)
MeSH only      12  0.3177  0.2014      (------*------)
GMed only      12  0.2737  0.2758    (------*-----)
Auto & MeSH    12  0.6889  0.3176                        (-----*------)
Auto & GMed    12  0.6864  0.1842                        (-----*------)
MeSH & GMed    12  0.5856  0.1976                   (-----*------)
                                    ----+---------+---------+---------+--
Pooled StDev =  0.2344                 0.20      0.40      0.60      0.80
```

Fig. 6. Recall comparison by term source.



```
ONE-WAY ANALYSIS OF VARIANCE FOR TERM PRECISION

Analysis of Variance
Source      DF     SS       MS        F       p
Precision    6   0.0491   0.0082    0.09   0.997
Error       77   6.6807   0.0868
Total       83   6.7298

                         Individual 95% CIs For Mean
                         Based on Pooled StDev
Level          N    Mean    StDev    ----+---------+---------+---------+--
Auto only      12   0.4012  0.2953       (-------------*--------------)
MeSH only      12   0.3723  0.3025    (-------------*-------------)
GMed only      12   0.3704  0.3948   (-------------*-------------)
Auto & MeSH    12   0.3639  0.2832  (-------------*-------------)
Auto & GMed    12   0.4365  0.2819        (-------------*-------------)
MeSH & GMed    12   0.3686  0.2517   (-------------*-------------)
All            12   0.3955  0.2229      (-------------*-------------)
                                     ----+---------+---------+---------+--
Pooled StDev =  0.2946                  0.24      0.36      0.48      0.60
```

Fig. 7. Precision comparison by term source.

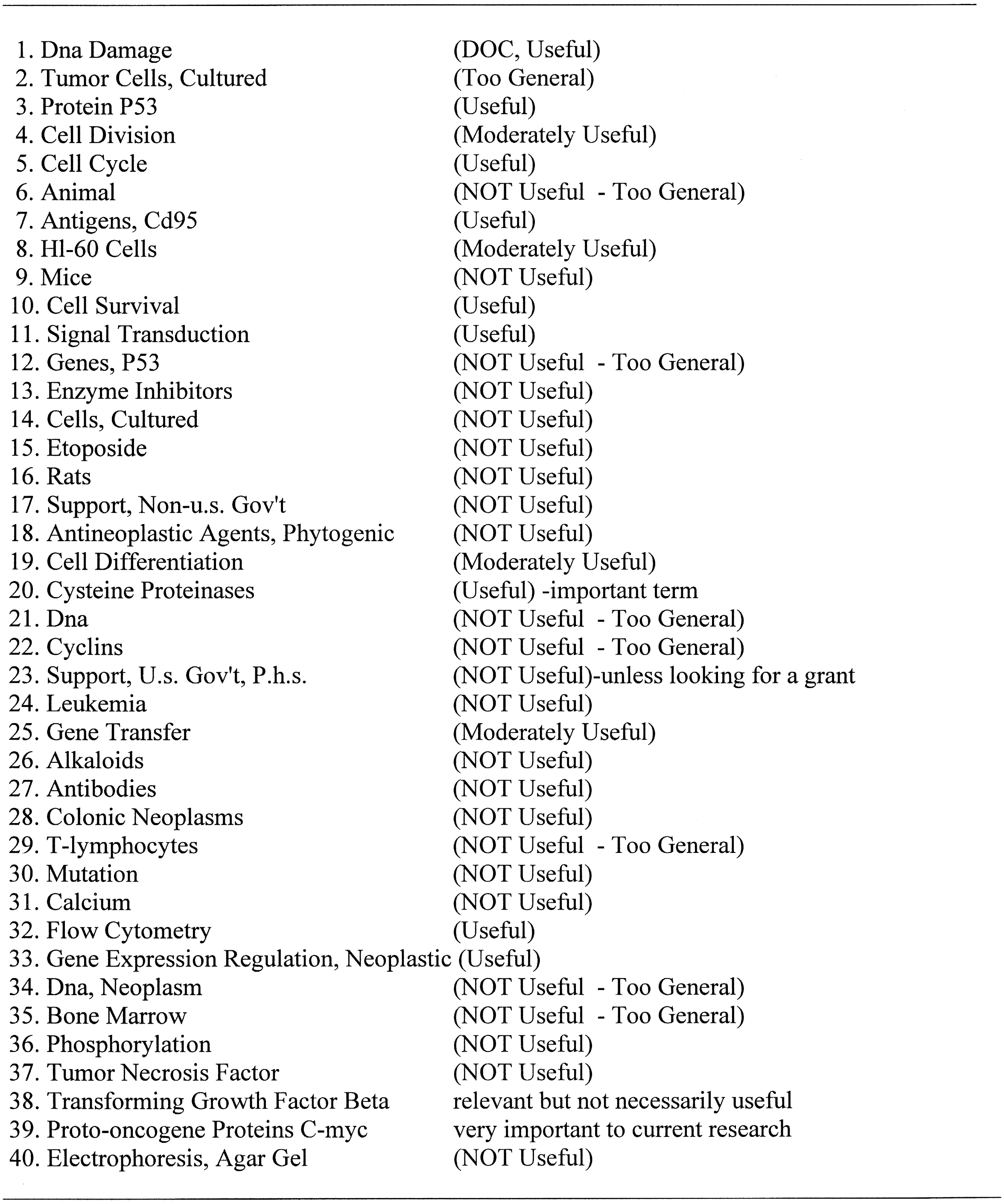
In general, our subjects liked the Automatic Indexing concept space best. They subjectively felt that it came up with the most interesting and most relevant terms the majority of the time. Instances when it did not (i.e. “Wiskott–Aldrich syndrome”) were explained (by the subjects) as follows: “That is a very specific and narrow topic. It is likely that it wasn’t mentioned in just 2 months of CANCERLIT abstracts, which is why your system can’t find it.” Subjects were very impressed by the quality of our concept space based on only 2 months of data, and most of them requested that we contact them when the larger collection’s concept spaces become available.

3.4. *Qualitative evaluation*

Figs. 8–10 illustrate a search using the subject-supplied term “apoptosis” (a type of cell death) for each of the three thesauri. The MeSH thesaurus suggested 40 related terms, of which, 10 terms were considered useful (two extremely useful), five moderately useful, and 25 not useful (eight were too general). The Automatic Indexing thesaurus also suggested 40 terms, of which 26 terms were rated useful (three extremely useful), three moderately useful, and 10 not useful (five were too general). Internet Grateful Med suggested nine terms (one useful, one moderately useful and seven not useful). There were three duplicate terms (all duplication occurred between the MeSH and Automatic Indexing thesauri).

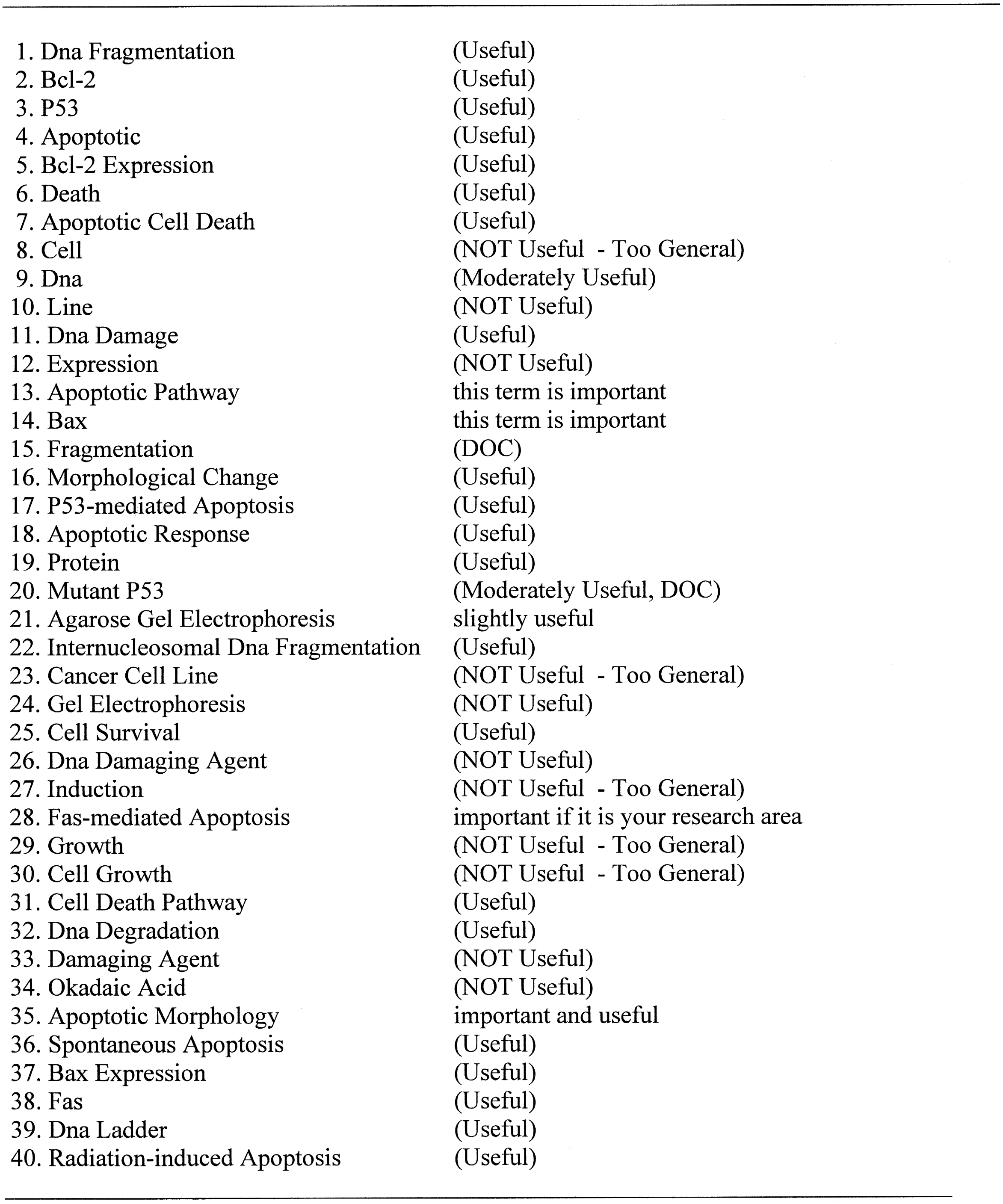
Most of our subjects were familiar with MeSH terms and some had previously used Internet Grateful Med or MEDLINE. Currently, most of them use OVID for their reference searching. One subject suggested extending the concept space to include all of MEDLINE instead of restricting it to CANCERLIT. Another subject, who had spent time at NIH, and had extensive experience with both Internet Grateful Med and MeSH terms, suggested that a MeSH-based thesaurus/Automatic Indexing concept space combination would be more effective. We showed him a combined MeSH/Automatic Indexing concept space and told him that future plans include incorporating the UMLS Metathesaurus.

Interestingly, we had difficulty getting subjects to suggest five *specific* relevant terms before the search



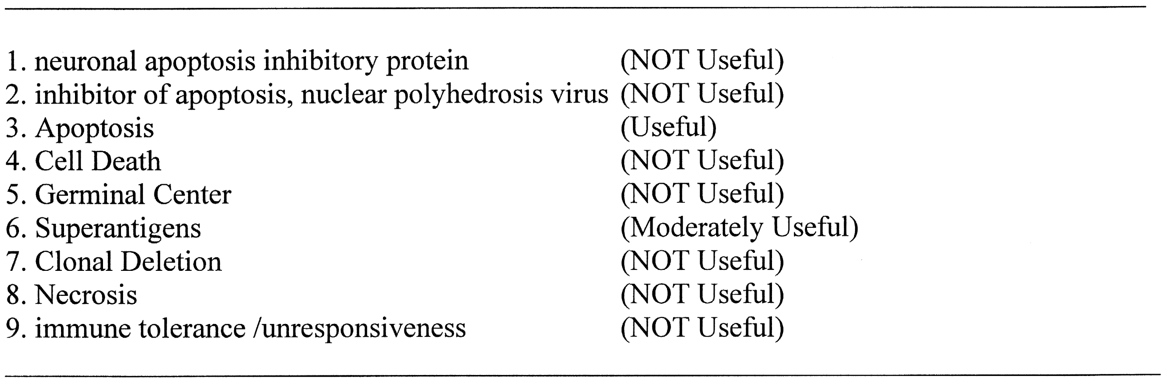
1. Dna Damage	(DOC, Useful)
2. Tumor Cells, Cultured	(Too General)
3. Protein P53	(Useful)
4. Cell Division	(Moderately Useful)
5. Cell Cycle	(Useful)
6. Animal	(NOT Useful - Too General)
7. Antigens, Cd95	(Useful)
8. Hl-60 Cells	(Moderately Useful)
9. Mice	(NOT Useful)
10. Cell Survival	(Useful)
11. Signal Transduction	(Useful)
12. Genes, P53	(NOT Useful - Too General)
13. Enzyme Inhibitors	(NOT Useful)
14. Cells, Cultured	(NOT Useful)
15. Etoposide	(NOT Useful)
16. Rats	(NOT Useful)
17. Support, Non-u.s. Gov't	(NOT Useful)
18. Antineoplastic Agents, Phytogenic	(NOT Useful)
19. Cell Differentiation	(Moderately Useful)
20. Cysteine Proteinases	(Useful) -important term
21. Dna	(NOT Useful - Too General)
22. Cyclins	(NOT Useful - Too General)
23. Support, U.s. Gov't, P.h.s.	(NOT Useful)-unless looking for a grant
24. Leukemia	(NOT Useful)
25. Gene Transfer	(Moderately Useful)
26. Alkaloids	(NOT Useful)
27. Antibodies	(NOT Useful)
28. Colonic Neoplasms	(NOT Useful)
29. T-lymphocytes	(NOT Useful - Too General)
30. Mutation	(NOT Useful)
31. Calcium	(NOT Useful)
32. Flow Cytometry	(Useful)
33. Gene Expression Regulation, Neoplastic	(Useful)
34. Dna, Neoplasm	(NOT Useful - Too General)
35. Bone Marrow	(NOT Useful - Too General)
36. Phosphorylation	(NOT Useful)
37. Tumor Necrosis Factor	(NOT Useful)
38. Transforming Growth Factor Beta	relevant but not necessarily useful
39. Proto-oncogene Proteins C-myc	very important to current research
40. Electrophoresis, Agar Gel	(NOT Useful)

Fig. 8. MeSH only concept space: terms related to apoptosis.



1. Dna Fragmentation	(Useful)
2. Bcl-2	(Useful)
3. P53	(Useful)
4. Apoptotic	(Useful)
5. Bcl-2 Expression	(Useful)
6. Death	(Useful)
7. Apoptotic Cell Death	(Useful)
8. Cell	(NOT Useful - Too General)
9. Dna	(Moderately Useful)
10. Line	(NOT Useful)
11. Dna Damage	(Useful)
12. Expression	(NOT Useful)
13. Apoptotic Pathway	this term is important
14. Bax	this term is important
15. Fragmentation	(DOC)
16. Morphological Change	(Useful)
17. P53-mediated Apoptosis	(Useful)
18. Apoptotic Response	(Useful)
19. Protein	(Useful)
20. Mutant P53	(Moderately Useful, DOC)
21. Agarose Gel Electrophoresis	slightly useful
22. Internucleosomal Dna Fragmentation	(Useful)
23. Cancer Cell Line	(NOT Useful - Too General)
24. Gel Electrophoresis	(NOT Useful)
25. Cell Survival	(Useful)
26. Dna Damaging Agent	(NOT Useful)
27. Induction	(NOT Useful - Too General)
28. Fas-mediated Apoptosis	important if it is your research area
29. Growth	(NOT Useful - Too General)
30. Cell Growth	(NOT Useful - Too General)
31. Cell Death Pathway	(Useful)
32. Dna Degradation	(Useful)
33. Damaging Agent	(NOT Useful)
34. Okadaic Acid	(NOT Useful)
35. Apoptotic Morphology	important and useful
36. Spontaneous Apoptosis	(Useful)
37. Bax Expression	(Useful)
38. Fas	(Useful)
39. Dna Ladder	(Useful)
40. Radiation-induced Apoptosis	(Useful)

Fig. 9. Automatic Indexing only concept space: terms related to apoptosis.



1. neuronal apoptosis inhibitory protein	(NOT Useful)
2. inhibitor of apoptosis, nuclear polyhedrosis virus	(NOT Useful)
3. Apoptosis	(Useful)
4. Cell Death	(NOT Useful)
5. Germinal Center	(NOT Useful)
6. Superantigens	(Moderately Useful)
7. Clonal Deletion	(NOT Useful)
8. Necrosis	(NOT Useful)
9. immune tolerance /unresponsiveness	(NOT Useful)

Fig. 10. Internet Grateful Med: terms related to apoptosis.

process began. Subjects were more comfortable suggesting *categories* of information (e.g. related drugs, treatment regimes) as opposed to specific terms (e.g. a specific drug or treatment). Later, during thesauri-suggested term evaluation, subjects frequently said, “That term is not identical to the one that I suggested, but it means the same thing.” (Nicely illustrating synonymy).

4. Conclusions and future directions

Different users with different goals approach large information spaces in different ways. We focused on medical researchers and a highly technical, research-based biomedical document collection (CANCERLIT). This type of medical information user is a very technical, extremely focused expert who is intimately familiar with a particular section of the information space. Our subjects were interested in very narrow, directed searches. Due to their busy schedules, they had no interest in browsing or exploring the collection. Based on their qualitative feedback, an automated thesaurus or concept space approach to indexing the CANCERLIT collection was preferred for information retrieval over the use of currently existing biomedical thesauri. We feel that this result is consistent with other research on concept space use in information retrieval.

We were especially encouraged by the precision and recall performance of the Automatic Indexing thesaurus (no statistical differences between it and the other two existing biomedical thesauri), since it was based on a very limited number of CANCERLIT documents (only 2 months’ worth of data). We believe it will be significantly better when a larger set of documents serves as its basis.

For this type of user, already familiar with biomedical terminology, a combined concept space that augments automatic indexing terms with terms from existing biomedical thesauri could potentially improve information retrieval. To this end, we are in the process of creating a set of concept spaces for the CANCERLIT collection that include MeSH terms and UMLS terms. Future plans may include incorporating the UMLS Semantic Network. An important advantage of including the UMLS information is that we may be able to use it to address the generalization/specialization criticism of statistical techniques. Statistical techniques do not take into account the term’s part of speech or level of abstraction because terms are analyzed statistically and syntactically, not semantically. The UMLS products capture a parent/child relationship between concepts and we may be able to use this feature to generalize and to organize terms by level of abstraction.

Other important future enhancements would be to allow the searcher to select what concept space to use and to allow a searcher to dynamically add terms of interest to the thesaurus for future indexing and retrieval. Ideally, future interfaces will allow subjects to interactively weigh both individual terms and term source to improve their searches.

Another common criticism of the concept space technique is that because it is *syntactic*, not semantic, it ‘analyzes’ terms “out of context”. To address this concern we are investigating incorporating a Natural Language parser at the front-end of our concept space analysis, allowing term analysis in the context of the noun or verb phrase in which they occur. We are currently investigating only noun phrase parsing. Qualitative feedback indicated that precise terms were especially important to many medical information users including primary care professionals (e.g. doctors, nurses) and medical research specialists. We believe that the precision and quality of our terms can be improved using Natural Language Processing techniques, which identify key noun phrases in documents. An Arizona Noun Phraser has been developed and implemented against both of our biomedical document testbeds (TOXLINE from NLM and CANCERLIT from NCI). Future research will involve investigating the impacts of Arizona Noun Phraser usage on usability and information retrieval quality.

Novice users and others unfamiliar with the CANCERLIT collection and/or biomedical terms may prefer to browse or explore as opposed to performing narrow directed searches. It is likely that this type of user will prefer other types of tools (for example, a Kohonen-based category map) over concept spaces and existing biomedical thesauri. Future research with the CANCERLIT collection will need to include a larger and more varied group of subjects and information retrieval tools.

Finally, medical information already contains both static and moving images. Several image indexing and retrieval techniques have been applied to medical image databases. Indeed any *complete* medical informatics system must address *image* indexing and retrieval, which are especially important to people who use patient record medical information. Our lab is currently investigating image indexing and retrieval (using visual thesauri and visual SOMs) and image similarity analysis on a GIS collection. Future research on medical information retrieval, in particular, patient record medical information, will include combining image indexing and textual information indexing and retrieval techniques.

The results from our current experimentation on the CANCERLIT and TOXLINE testbeds are preliminary, but encouraging. In our ongoing effort in the Illinois Digital Library Initiative project, we are in the process of fine-tuning these techniques and exploring other general-purpose artificial intelligence and mathematical pattern analysis techniques for various digital library and medical information retrieval and analysis applications.

Acknowledgements

This project was funded primarily by a grant from the NCI “Information Analysis and Visualization for Cancer Literature” (1996–1997), a grant from the NLM, “Semantic Retrieval for Toxicology and Hazardous Substance Databases” (1996–1997), an NSF/CISE “Intelligent Internet Categorization and Search” project (1995–1998), and the NSF/ARPA/NASA Illinois Digital Library Initiative project, “Building the Interspace” (1994–1998).

We would like to thank the following individuals for their generous donation of time and their thoughtful evaluations and suggestions: Dr. Margaret Briel, Dr. Sherri Chow, Dr. Bernie Futcher, Dr. Jesse Martinez, and Dr. Luke Whitecell.

The CANCERLIT collection is available through many sources including:

1. CANCERLIT [database online] Bethesda (MD): NCI, 1997. Available through the NLM, Bethesda, MD, and
2. CANCERLIT [Internet URL]www.ovid.com. Available through Ovid Technologies, New York, NY.

References

[1] B.T. Bartell, G.W. Cottrell, R.K. Belew, Representing documents using an explicit model of their similarities, Journal of the American Society for Information Science 46 (4) (1995) 254–271.

[2] M.J. Bates, Subject access in online catalogs: A design model, Journal of the American Society for Information Science 37 (6) (1986) 357–376, November.

[3] D.C. Blair, M.E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, Communications of the ACM 28 (3) (1985) 289–299.

[4] H. Chen, K.J. Lynch, Automatic construction of networks of concepts characterizing document databases, IEEE Transac-

tions of Systems, Man and Cybernetics 22 (5) (1992) 885–902, September/October.

[5] H. Chen, J. Martinez, D.T. Ng, B.R. Schatz, A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment of the Worm Community System, Journal of the American Society for Information Science 48 (1) (1997) 157–170, February.

[6] G.C. Chute, Y. Yang, D.A. Evans, Latent semantic indexing of medical diagnoses using UMLS semantic structures, Proceedings of the 15th SCAMC, Washington, DC, 1991, pp. 185–189.

[7] J.J. Cimino, S.B. Johnson, P. Peng, A. Aguirre, From ICD9-CM to MeSH using the UMLS: A how-to guide, In Proceedings — the Annual Symposium on Computer Applications in Medical Care, 1994, pp. 730–734.

[8] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407, September.

[9] F.C. Ekmekcioglu, A.M. Robertson, P. Willett, Effectiveness of query expansion in ranked-output document retrieval systems, Journal of Information Science 18 (1992) 139–147.

[10] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais, The vocabulary problem in human–system communication, Communications of the ACM 30 (11) (1987) 964–971, November.

[11] P.N. Gorman, Information needs of physicians, Journal of the American Society for Information Science 46 (10) (1995) 729–736.

[12] J.N. Guidi, E.A. Fox, Information retrieval and genomics — An introduction, Computers in Biology and Medicine 26 (3) (1996) 179, May.

[13] W.R. Hersh, The electronic medical record: Promises and problems, Journal of the American Society for Information Science 46 (10) (1995) 772–776.

[14] W.R. Hersh, Information Retrieval: A Health Care Perspective, Springer, New York, NY, 1995.

[15] S.M. Huff, J.J. Cimino, Medical data dictionaries and their use in medical information system development, in: U. Prokosch, J. Dudeck (Eds.), Hospital Information Systems: Design and Development Characteristics; Impact and Future Architecture, Elsevier, 1995, pp. 53–75.

[16] D.A.B. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods of Information in Medicine 32 (1993) 281–291.

[17] S. Nadis, Computation cracks semantic barrier between databases, Science 272 (1996) 1419, 7 June.

[18] P.W. Richwine, R. Lilly, A study of MeSH and UMLS for subject searching in an online catalog, Bulletin of the Medical Library Association 81 (2) (1993) 229–233, April.

[19] G. Salton, Automatic Text Processing, Addison-Wesley Publishing, Reading, MA, 1989.

[20] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science 41 (4) (1990) 288–297.

[21] P. Srinivasan, Optimal document-indexing vocabulary for MEDLINE, Information Processing & Management 32 (4) (1996) 431–443.

Dr. Andrea Houston is an Assistant Professor in the Information Systems and Decision Sciences Department at Louisiana State University. She received her PhD from the Department of Management Information Systems at The University of Arizona in 1998, her MBA from the University of New Hampshire (1981) and her BA (1976) from the University of Pennsylvania. She worked in industry as a project leader and systems analyst for over 15 years before returning to academics. Her research interests include looking at information retrieval issues, medical informatics, digital libraries and electronic publishing, human factors in Human/Computer Interaction, and Natural Language Processing. She is also interested in software team development and organizational memory. She is a member of ACM, IEEE, AIS and ASIS.

Dr. Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona and head of the UA/MIS Artificial Intelligence Lab. He received the PhD degree in Information Systems from New York University in 1989. He is author of more than 70 articles covering semantic retrieval, search algorithms, knowledge discovery, and collaborative computing in leading information technology publications. He serves on the editorial board of Journal of the American Society for Information Science and Decision Support Systems. He is an expert in digital library and knowledge management research whose work has been featured in various scientific and information technologies publications including Science, Business Week, NCSA Access Magazine, WEBster, and HPCWire.

Susan M Hubbard, RN, is the director of the International Cancer Information Center (since 1984) and an Associate Director of the National Cancer Institute. She received a BS from the Honors College of the University of Connecticut and a Master's in Public Administration from the American University (1993). As the Director of NCI's International Cancer Information Center, she directs the NCI's efforts to identify, implement, and evaluate state-of-the-art communication technologies to maximize the potential impact of advances in cancer research on health care. She directed the development of PDQ, NCI's primary mechanism for communicating state-of-the-art information about cancer. The Department of Health and Human Services (DHHS) designated ICIC as a "reinvention laboratory" in 1994 under Vice President Albert Gore's National Performance Review (NPR). The International Cancer Information Center has also received the NPR "Hammer Award" and the Department of Health and Human Service's Continuous Improvement Program Award. Ms. Hubbard has published extensively in nursing and medical journals and texts.

Bruce R. Schatz is Director of the Community Architecture for Network Information SYSTEMS (CANIS) Laboratory at the University of Illinois at Urbana–Champaign. He is the Principal Investigator of the Digital Libraries Initiative project and the DARPA Information Management Program, which performs research in information systems building analysis environments to support community repositories (Interspace), and in information science performing large-scale experiments in semantic retrieval for vocabulary switching. Dr. Schatz holds faculty appointments in Library and Information Science, Computer Science, Neuroscience, and Health Information Sciences. He is also a Senior Research Scientist at the National Center for Supercomputing Applications (NCSA), serving as the scientific advisor for digital libraries and information systems. He has served in this role since 1989, including the period during which NCSA developed Mosaic.

Robin Sewell received her Doctor of Veterinary Medicine degree from Washington State University (1986) and her MLA from the University of Arizona (1997). She currently serves as the AI Lab's Program Coordinator and a Research Specialist. Her interests are in Medical Informatics and the National Library of Medicine's Unified Medical Language System.

Dorbin Ng received the PhD in 2000 from the Department of Management Information Systems at the University of Arizona, from which he also received a BS in Business Administration majoring in Management Information Systems and Finance (1990) and a MS in MIS (1993). He is currently a Systems Scientist in the Computer Science Department at Carnegie Mellon University working on the Informedia Digital Video Library Project. His research interests include digital libraries, intelligence and multimedia information retrieval, semantic interoperability for information analysis and knowledge management environment, large-scale knowledge discovery using high-performance supercomputers, search engine and user interface development in Internet, neural-network computing, and collaborative computing.